SeqTrace: A Graphical Tool for Rapidly Processing DNA Sequencing Chromatograms

*Brian J. Stucky**

Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA

Modern applications of Sanger DNA sequencing often require converting a large number of chromatogram trace files into high-quality DNA sequences for downstream analyses. Relatively few nonproprietary software tools are available to assist with this process. SeqTrace is a new, free, and open-source software application that is designed to automate the entire workflow by facilitating easy batch processing of large numbers of trace files. SeqTrace can identify, align, and compute consensus sequences from matching forward and reverse traces, filter low-quality base calls, and end-trim finished sequences. The software features a graphical interface that includes a full-featured chromatogram viewer and sequence editor. SeqTrace runs on most popular operating systems and is freely available, along with supporting documentation, at http://seqtrace.googlecode.com/.

Key Words: bioinformatics, sequence analysis, software

INTRODUCTION

Since its development in the late 1970s, Sanger chain-termination DNA sequencing[1] has become a widely used, essential technique of molecular biology.[2] Although high-coverage, high-volume sequencing has largely moved to "next-generation" technologies, Sanger sequencing remains a popular and indispensable tool for low-coverage sequencing applications, such as phylogenetic analyses or DNA barcoding efforts.[3,4]

Many such projects require high-quality sequencing reads from a relatively large number of PCR amplicons. Modern Sanger sequencing instruments, however, generate "raw" chromatogram trace files that require further processing to obtain sequences of sufficient quality for downstream analyses. At a minimum, this involves inspecting each trace file to identify problematic sequencing runs, remove unreliable base calls, and trim the ends of the sequence. Paired forward and reverse reads of PCR products are also frequently used to ensure final sequence quality; this requires aligning the forward and reverse sequences and determining a single consensus sequence from the pair. If done manually, these steps can be very time consuming, especially for large projects. Although commercial software is available to handle these tasks, free software options are generally much more limited.

SeqTrace, a new computer program described in this communication, was created to help fill this gap. SeqTrace is intended specifically for sequencing projects that require converting trace files directly into high-quality, finished sequences, and it provides a graphical, user-friendly interface for automating the entire process. Although SeqTrace was designed with batch processing in mind, it can also serve as a general-purpose trace viewer and editor. SeqTrace is free and open-source software that runs on all popular operating systems.

MATERIALS AND METHODS

SeqTrace was designed to be a graphical, user-friendly software program that would run on most of the common operating systems in current use. A secondary goal was to ensure that the SeqTrace source code could be reused easily in other bioinformatics applications. To meet these requirements, SeqTrace follows object-oriented design principles and was implemented in Python (http://www.python.org/) using the cross-platform GTK+ windowing toolkit (http://www.gtk.org/). To support multiple input and ouput file formats without requiring the user to install additional software libraries, all file formats were implemented directly in Python as part of the SeqTrace application.

Generating a consensus sequence from matching forward and reverse sequencing reads requires first computing

*Address correspondence to: Brian J. Stucky, Dept. of Ecology and Evolutionary Biology, University of Colorado, Boulder, Ramaley N122, Campus Box 334, Boulder, CO 80309-0334, USA. (E-mail: stuckyb@colorado.edu).

I have no financial support or associations that pose a conflict of interest for the work reported in this manuscript.

doi: 10.7171/jbt.12-2303-004

a pairwise global alignment of the raw forward and reverse sequences. To accomplish this, SeqTrace uses a customized Needleman-Wunsch pairwise alignment algorithm.[5,6] Base mismatches, interior gap openings, and gap extensions are all given the same penalty (i.e., linear gap penalties are used). For matched forward and reverse sequences, base mismatches and gaps both represent sequencing errors and should be weighted equally.

Accurate base-call quality scores are essential for producing finished sequences. Calculating accurate quality scores is a complex problem, due in part to variations in sequencing machines and techniques.[7,8] However, all modern capillary sequencing instruments use base-calling software that calculate quality scores, so virtually all recent chromatogram trace files include them. Consequently, trace files processed with SeqTrace are expected to include quality scores, and SeqTrace does not attempt to compute them if they are absent.

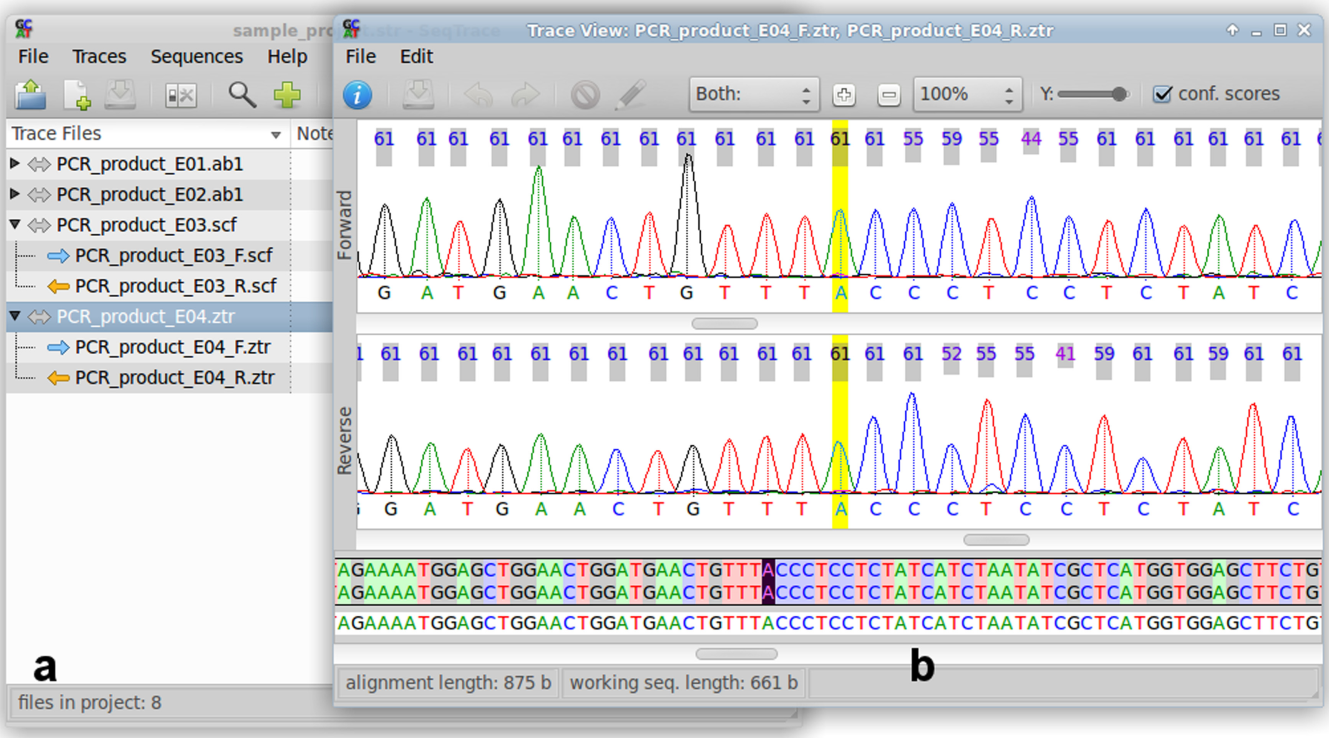
SeqTrace was tested on several popular GNU/Linux distributions (Xubuntu 11.10, Ubuntu 11.04, and Linux Mint 12), recent versions of Microsoft Windows (XP, Vista, and 7), and Apple OS X 10.6 in the native and X11 environments. Additionally, PyInstaller (http://www.pyinstaller.org/) was used to create a self-contained SeqTrace package for Microsoft Windows that eliminates all software dependencies on that platform.

RESULTS

The SeqTrace program, source code, and supporting documentation are all freely available at http://seqtrace.googlecode.com. SeqTrace is distributed as a binary package for Windows and as a source package for all other operating systems. SeqTrace has modest hardware requirements: computers capable of running modern graphical desktop operating systems should also be adequate for processing trace files with SeqTrace. The SeqTrace source code is licensed under version 3 of the GNU General Public License (http://www.gnu.org/licenses/gpl.html).

SeqTrace supports several input trace file formats, including Applied Biosystems, Inc., Format,[9] the machine-independent Standard Chromatogram Format,[10] and the highly space-efficient ZTR format.[11] SeqTrace includes the first open-source Python implementation of the ZTR file format, code which will likely be of use to other Python-based bioinformatics projects.

To work with sequencing trace files in SeqTrace, the user can either open individual trace files directly or create a SeqTrace project containing one or many trace files. SeqTrace projects provide a convenient way to organize large numbers of trace files and their associated DNA sequences. Projects are managed from the main SeqTrace window (Fig. 1A), where the user can add or remove files


sample_pro
File
Traces
Sequences
Help
Trace Files
Note
PCR_product_E01.ab1
PCR_product_E02.ab1
PCR_product_E03.scf
PCR_product_E03_F.scf
PCR_product_E03_R.scf
PCR_product_E04.ztr
PCR_product_E04_F.ztr
PCR_product_E04_R.ztr
files in project: 8
a
Trace View: PCR_product_E04_F.ztr, PCR_product_E04_R.ztr
File
Edit
Both:
100%
Y:
conf. scores
Forward
Reverse
G A T G A A C T G T T T A C C C T C C T C T A T C
AGAAAATGGAGCTGGAACTGGATGAACTGTTTACCCTCCTCTATCATCTAATATCGCTCATGGTGGAGCTTCTG
AGAAAATGGAGCTGGAACTGGATGAACTGTTTACCCTCCTCTATCATCTAATATCGCTCATGGTGGAGCTTCTG
AGAAAATGGAGCTGGAACTGGATGAACTGTTTACCCTCCTCTATCATCTAATATCGCTCATGGTGGAGCTTCTG
b

alignment length: 875 b
working seq. length: 661 b

FIGURE 1

SeqTrace's user interface, including the project window (*A*) and the trace-view window (*B*).

from the project, indicate whether files are forward or reverse sequencing reads, and group matching forward and reverse reads together. SeqTrace can save projects to an external file using a custom file format that preserves the state of the project and all of its sequences.

Sequencing trace files are examined in the trace-viewing window (Fig. 1B). Multiple trace files can be viewed simultaneously, each in their own window. The user has full control over zooming, scaling, and scrolling the rendered electropherograms. SeqTrace can use the quality scores associated with the base calls to identify ambiguous bases and trim regions of low quality from the beginning and end of the sequences. The user can also directly edit the sequence from within the viewing window. Editing includes unlimited undo and redo functionality.

Matching forward and reverse traces can be opened together in the trace-viewing window (Fig. 1B). SeqTrace will automatically reverse-complement the reverse trace and sequence, align the forward and reverse sequences, and then use the base-call quality scores to compute a single consensus sequence from the alignment. This consensus sequence can be custom edited by the user in the same manner as a sequence from a single trace file.

A major feature of SeqTrace is its ability to automate the processing of large numbers of chromatogram files. SeqTrace can batch-process all of the trace files in a project, resulting in a set of high-quality finished sequences. To accomplish this, SeqTrace first uses customizable search strings to automatically identify matching forward and reverse trace files. SeqTrace then aligns the matched forward and reverse traces, calculates their consensus sequences, identifies low-quality base calls, and quality-trims the ends of the consensus sequences. When batch-processing trace files, SeqTrace can easily handle hundreds of trace files in a single project in a matter of seconds.

Finished sequences can be exported singly from individual trace files or simultaneously from one or more trace files in a SeqTrace project. SeqTrace can also export matching forward and reverse sequence alignments. Supported export file formats include FASTA (http://fasta.bioch.virginia.edu), NEXUS,[12] or a simple text format.

DISCUSSION

SeqTrace is designed to streamline sequencing projects where each chromatogram trace file or pair of forward and reverse trace files must be converted into a single, high-quality DNA sequence. Thus, SeqTrace is well-suited for many current applications of Sanger sequencing, which often involve low-coverage sequencing of individual positions.[4] It is important to note, however, that SeqTrace is not intended for sequencing scenarios that require contig assembly. Other software is available for such projects, such as the open-source Gap5.[13]

A variety of no-cost, open-source, and proprietary software packages are available that provide at least some of the functionality of SeqTrace. Most of these are limited to viewing and editing single trace files and do not include any support for batch processing. The earliest freely available, noncommercial trace viewer and editor was ted,[14] which was later replaced by trev.[15] Since then, several alternative programs have been released, including 4Peaks (Mekentosj, Amsterdam), Chromas Lite (Technelysium Pty., Australia), Chromatogram Explorer Lite (Heracle BioSoft S.R.L., Romania), Chromaseq (http://mesquiteproject.org/packages/chromaseq), ChroView,[16] FinchTV (Geospiza, Seattle, WA, USA), TraceEdit,[17] and Unipro UGENE (http://ugene.unipro.ru). Of these, Chromaseq is most similar to SeqTrace in terms of functionality.

SeqTrace offers several advantages in comparison with these earlier programs. First, SeqTrace is free and open-source software, unlike most of the alternatives [trev, Chromaseq, and UGENE are also open-source, although Chromaseq requires the programs phrap and phred (http://www.phrap.org), both of which are proprietary]. Second, SeqTrace is a self-contained application that does not require the installation of any additional bioinformatics software packages. Finally, SeqTrace is the only stand-alone program that combines sophisticated handling of trace files, including automatic alignment and calculation of consensus sequences from matching forward and reverse reads, with the ability to easily automate the processing of large numbers of trace files. For researchers that need to derive finished sequences from many chromatogram files, SeqTrace can result in considerable time savings in comparison with most other free programs. These capabilities, along with a user-friendly and intuitive interface, make SeqTrace an efficient tool for generating high-quality DNA sequences from chromatogram trace files.

ACKNOWLEDGMENTS

Rob Guralnick provided valuable feedback during the development of the software and on an early version of the manuscript. Erin Stucky also provided helpful comments that improved the manuscript. Michael Zianni kindly provided sequencing technology usage data and helped track down several bugs in the software. Finally, I thank three anonymous reviewers for their thoughtful critiques of the manuscript and software, both of which were improved as a result.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;74:5463–5467.
2. Pevsner J. *Bioinformatics and Functional Genomics*, 2nd ed. Hoboken, NJ, USA: Wiley-Blackwell, 2009;1–951.
3. Kieleczawa J, Adam D, Schweitzer P, Vennemeyer E, Zianni M.

Current state and future of capillary electrophoresis and Sanger sequencing. *J Biomol Tech* 2011;22:S14–S15.

4. Kircher M, Kelso J. High-throughput DNA sequencing—concepts and limitations. *BioEssays* 2010;32:524–536.
5. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;48:443–453.
6. Sankoff D. Matching sequences under deletion/insertion constraints. *Proc Natl Acad Sci USA* 1972;69:4–6.
7. Ewing B, Green P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* 1998;8:186–194.
8. Walther D, Bartha G, Morris M. Basecalling with LifeTrace. *Genome Res* 2001;11:875–888.
9. Life Technologies Corporation. *Applied Biosystems Genetic Analysis Data File Format*. 2009;1–56.
10. Dear S, Staden R. A standard file format for data from DNA sequencing instruments. *DNA Seq* 1992;3:107–110.
11. Bonfield JK, Staden R. ZTR: a new format for DNA sequence trace data. *Bioinformatics* 2002;18:3–10.
12. Maddison DR, Swofford DL, Maddison WP. NEXUS: an extensible file format for systematic information. *Syst Biol* 1997; 46:590–621.
13. Bonfield JK, Whitwham A. Gap5—editing the billion fragment sequence assembly. *Bioinformatics* 2010;26:1699–1703.
14. Gleeson T, Hillier L. A trace display and editing program for data from fluorescence based sequencing machines. *Nucleic Acids Res* 1991;19:6481–6483.
15. Bonfield JK, Beal KF, Betts MJ, Staden R. Trev: a DNA trace editor and viewer. *Bioinformatics* 2002;18:194–195.
16. Tae H, Kong E-B, Park K. ChroView: a trace viewer for browsing and editing chromatogram files. *Genomics Inform* 2007;5:30–31.
17. Rothgänger J, Weniger M, Weniger T, Mellmann A, Harmsen D. Ridom TraceEdit: a DNA trace editor and viewer. *Bioinformatics* 2006;22:493–494.